TOBB ETU at CheckThat! 2023: Utilizing ChatGPT to Detect Subjective Statements and Political Bias

Notebook for the CheckThat! Lab at CLEF 2023

Mehmet Deniz Türkmen, Gökalp Coşgun and Mucahid Kutlu

TOBB University of Economics and Technology, Ankara, Turkey

Abstract

Information has been referred to as the "oil" of the 21st century, emphasizing its immense importance. However, it also has the potential to pose significant risks and hazards if it is not correct. Hence, it is imperative to reduce the spread of misinformation. In this paper, we present our participation in Task 2 (i.e., detecting subjective tasks) and Task 3A (i.e., detecting political bias in news articles) of CLEF CheckThat! 2023 which focuses on reducing the spread of misinformation. We propose utilizing ChatGPT for these classification tasks and explore zero-shot and few-shot classification using ChatGPT. While the performance of our approach varies across different languages in Task 2, we are ranked 3rd on the German dataset with 0.71 macro F_1 score. In Task 3A, we are ranked 2nd with 0.646 Mean Absolute Error (MAE).

Keywords

fact-checking, subjectivity, political bias, shared task

1. Introduction

The Internet has served as the primary tool for accessing information over the past two decades. Alongside its ability to facilitate the rapid dissemination of information, it provides a wide range of information sources thanks to its inclusive nature, allowing contributions from anyone. While this accessibility makes reaching information extremely convenient, it also raises concerns about the reliability and quality of the information available. As a result, ensuring the trustworthiness of information has become an important research direction [1, 2].

The process of verifying factuality is not a straightforward task and often requires multiple steps to reach a conclusive result. One crucial step in fact-checking is to decide whether a statement requires fact-checking or not. For instance, we first need to detect the subjectivity of a statement as there is no need to fact-check personal opinions and beliefs. In addition, when content creators exhibit personal biases, the reliability of the information necessitates further investigation. In this regard, we focus on subjectivity and

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18-21, 2023, Thessaloniki, Greece

✉ m.turkmen@etu.edu.tr (M. D. Türkmen); gcosgun@etu.edu.tr (G. Cogun); mkutlu@etu.edu.tr (M. Kutlu)

political bias analysis, specifically addressing Task 2 [3] and Task 3A [4] in CheckThat! Lab at CLEF 2023 [5].

Subjectivity analysis entails identifying whether a sentence is subjective or objective, while political bias analysis involves detecting left, right, and center opinions within an article. In our study, we propose the utilization of ChatGPT for both tasks. Renowned for generating high-quality responses across various domains, we think that it can be also utilized for classification problems, as it eliminates the need for training or tuning processes. Thus, we adapt ChatGPT to the classification tasks in CheckThat! 2023 Lab and evaluate its performance in zero-shot and few-shot settings.

Our findings can be summarised as follows: 1) With the exception of the Dutch and German datasets, ChatGPT falls behind the baseline method in Task 2. 2) ChatGPT is more effective in detecting the political bias than subjectivity. Our approach in Task 3A is ranked 2^{nd}. 3) Interestingly, ChatGPT performs better in zero-shot classification compared to the few-shot setting in both tasks.

2. Related Work

2.1. Subjectivity Detection

Subjectivity analysis has been extensively explored by prior work. Numerous studies have dedicated efforts towards subjectivity detection, recognizing its importance as a preliminary step in other NLP tasks [6, 7]. Chaturvedi et al. [7] and Montoyo et al. [8] specifically concentrate on subjectivity detection to enhance the quality of sentiment analysis. In the case of Sixto et al. [6], subjectivity analysis serves as the primary step in polarity detection tasks. Riloff [9] exploits subjectivity to enhance the precision of information extraction systems. Wilson et al. [10] design a comprehensive system for detailed subjectivity analysis. Apart from subjective text classification, their system also aims to extract textual elements that contribute to subjectivity.

2.2. Political Bias

An extensive amount of research has been conducted to investigate the presence of political bias in news articles, shedding light on its impact and various dimensions [11, 12]. Content based bias detection is generally conducted at two levels of granularity: the article level [13] and the sentence level [14]. Lin et al. [15] use statistical methods such as Naive Bayes and SVM to analyze political orientation at the document and sentence levels as a precursor to this problem. Chen et al. [12] demonstrate the detection of media bias using sequential models and illustrate the possibility to reveal the bias at different granularity levels. More recently, Hong et al. [11] propose a more robust and general multi-head attention by overcoming the issue of domain dependency.

3. Proposed Approach

In this section, we describe how we employ ChatGPT for zero-shot and few-shot classification. Firstly, we convert classification tasks into queries in order to interact with ChatGPT. These queries comprise two components: a command part and a content part. While the content part corresponds to the text to be classified, the command part specifies the desired output based on the content. In our scenario, the content corresponds to a sentence (Task 2) or a news article (Task 3A). The command presents labels and instructs ChatGPT to classify the content based on the labels. We use separate commands for the two tasks we participate in.

Table 1
Zero-shot and few-shot subjectivity classification using ChatGPT for Task 2

Task and Method	Query	ChatGPT Response
Task 2, Zero-shot	Can you tell in one word whether this article expresses an objective or subjective opinion: ”It’s all justified in the name of racial “equity. But it’s nothing more than a nice way of saying they want to destroy the Historic American Nation.”	Subjective
Task 2, Few-shot	Can you tell in one word whether this article expresses an objective or subjective opinion: ”It’s all justified in the name of racial “equity. But it’s nothing more than a nice way of saying they want to destroy the Historic American Nation.” Here are some examples: 1) Theresa May made this more explicit: ‘Socialism is about levelling down. (Objective) 2) But he said the risk of a violent backlash had grown this year. (Objective) 3) Many originate with educational, recreational and sociological enthusiasts [...]. (Objective) 4) This week, authorities in Belgrade put a stop to EuroPride, [...]. (Objective) 5) “Normally, the majority opinion would speak for itself. The decision is [...]. (Objective) 6) White House officials have touted their efforts to cut down on the paperwork [...]. (Subjective) 7) One day’s work in every four belongs to government. (Subjective) 8) The battle to set our economic machine in motion in this emergency takes new [...]. (Subjective) 9) An indebted state-owned bus company in Lanzhou, the capital of Gansu [...] (Subjective) 10) Still, many of the nation’s 3.1 million public-school teachers have become [...]. (Subjective)	Subjective

🟧 Command 🟩 Content 🟪 Groundtruth examples

In the context of few-shot classification, our objective is to boost classification performance by providing ChatGPT with groundtruth classification examples, which consist of pairs of statements and their corresponding labels. To determine which examples to provide in Task 2, we initially conduct zero-shot classification on the training data. Subsequently, we randomly choose five misclassified training samples from each class and utilize them as ground-truth examples for few-shot classification on the validation and test data. In Task 3A, we follow a slightly different approach such that we randomly select five examples from the whole training set (i.e., not only misclassifed ones) for each label and provide these samples for few-shot classification. For each prediction, we re-sample the examples. Table 1 and Table 2 illustrate how we utilize ChatGPT for zero-shot and few-shot classification.

One of the problems we face during the utilization of ChatGPT for classification is that it cannot draw a conclusion about the class of some data samples. ChatGPT couldn't classify only a few of the samples in the dataset of Task 2, for each language. Approximately 100 data points for Task 3A responded to this text is not political, remaining neutral. In such cases, we naively classify those samples as the majority class, which is "OBJ" for Task 2 and "right" for Task 3A.

In Task 3A, we encountered some further challenges. We normally use the content part articles in our queries. However, the content part of 109 articles (out of 5008) contain the same text with different titles. We encountered the same issue with 45 samples in the test set. For these articles, we combine the title and content sections in our queries. In addition, in some cases ChatGPT returns "right-center" and "left-center" responses while we do not have these labels. In these cases, we select the majority class in the train-validation data.

Table 2
Zero-shot and few-shot political bias prediciton using ChatGPT for Task 3A

Task and Method	Query	ChatGPT Response
Task 3, Zero-shot	Choose a political leaning for this text, answer with only one word(left,center,right): "After 2½ years of civil war in Syria, President Barack Obama's larger policy is in disarray even as his administration, with help from Russia, averted a military..."	Center
Task 3, Few-shot	Choose a political leaning for this text, answer with only one word(left,center,right): "After 2½ years of civil war in Syria, President Barack Obama's larger policy is in disarray even as his administration, with help from Russia, averted a military..." Here are some examples of the classes: 1 - "Senate Judiciary Committee Chairman Lindsey Graham (R-S.C.) said [...]." : (Center) 2 - "The Senate confirmed Antony Blinken to be President Biden's Secretary of State [...]." : (Center) 3 - "President Biden directed the Department of Energy on Tuesday to release 50 [...]." : (Center) 4 - "News 'Every Blessing To Her And Her Film': Shia LaBeouf Reacts After Olivia [...]." : (Center) 5 - "News What to expect as Democrats retain the Senate for the next two years [...]." : (Center)' 6 - "As the House of Representatives is dragged closer to a vote on authorizing [...]." :(Left) 7 - "ANALYSIS Democrats came into the 2020 Senate elections as slight favorites [...]." : (Left) 8 - "News Iranian Women Are Burning Their Hijabs And Cutting Their Hair [...]." : (Left) 9 - "The emails to H don't contain a smoking gun, at least not yet [...]." : (Left) 10 - "ANALYSIS The growing Trump-Biden war over China, explained [...]." : (Left) 11 - "Yesterday during a press conference, Attorney General Eric Holder [...]." : (Right) 12 - "He was not on the ballot, but former President Trump was one of [...]." : (Right) 13 - "The scandals facing the White House particularly the Benghazi [...]." : (Right) 14 - "News Hearing aids now available over the counter for first time [...]." : (Right) 15 - "The Supreme Court delivered a dramatic change to abortion jurisprudence [...]." : (Right)	Center

🟧 Command 🟩 Content ⬜ Groundtruth examples

4. Experiments

In this section, we first describe our experimental setup. Subsequently, we present our results.

Table 3
Label distribution of train, validation and test sets for Task 2

	Train		Validation		Test	
Language	**OBJ**	**SUBJ**	**OBJ**	**SUBJ**	**OBJ**	**SUBJ**
Multilingual	4371	2257	300	300	300	300
Arabic	905	280	227	70	363	82
Dutch	489	311	107	93	263	237
English	353	298	106	113	116	127
German	492	308	123	77	194	97
Italian	1231	382	167	60	323	117
Turkish	422	378	100	100	129	111

Table 4
Label distribution of train, validation and test sets for Task 3A

Train			Validation			Test		
Left	**Center**	**Right**	**Left**	**Center**	**Right**	**Left**	**Center**	**Right**
12074	15452	17545	1342	1717	1949	2589	1959	650

4.1. Experimental Setup

The dataset created for Task 2 comprises 6 languages and the multilingual setting. Table 3 presents language-specific label distribution. For Task 3A, the training and development datasets consist of JSON files that contain title, content and label data. The data distributions of the train and validation sets are presented in Table 4. Due to time limitation and slow execution of ChatGPT results, we could use only 1,000 samples from the validation set in Task 3A. We use Turkish queries for Turkish dataset and English queries for the other languages.

In Task 2, we report accuracy and macro F_1 scores. In Task 3A, we report mean absolute error (MAE), i.e., the official metric of the lab, and accuracy, precision, recall, and f_1. Due to the imbalance label distribution, we present the weighted scores for all metrics except MAE.

4.2. Experimental Results

4.2.1. Task 2

Table 5 presents zero-shot and few-shot classification performances on the validation dataset. The evaluation of few-shot classification does not include multilingual data, mainly because the data samples which can belong to any of the six languages makes it challenging to select the appropriate ground-truth examples. As can be seen in the table, the zero-shot classification outperforms the few-shot classification, except for Italian and Turkish. Thus, we submitted few-shot predictions for the test data in these two languages, while utilizing zero-shot predictions for the others.

Table 6 shows the scores and rankings obtained from the submitted predictions for each language on the test dataset. In comparison to the validation scores, the test scores

Table 5
Zero-shot and few-shot classification results for Task 2 validation set

	Zero-shot		Few-shot	
	Accuracy	Macro F1	Accuracy	Macro F1
Multilingual	**0.69**	**0.68**	-	-
Arabic	**0.76**	**0.70**	0.51	0.51
Dutch	**0.64**	**0.64**	0.61	0.59
English	**0.68**	**0.66**	0.64	0.60
German	**0.79**	**0.78**	0.76	0.75
Italian	0.63	0.62	**0.73**	**0.69**
Turkish	0.73	0.72	**0.80**	**0.79**

are generally lower, with the exception of Dutch. We achieve the best ranking in German. While the subjectivity detection performance of ChatGPT is impressive, its performance is lower than most of the participants.

Table 6
Zero-shot and few-shot classification results for Task 2 test set

	Zero-shot		Few-shot		Rank	Baseline Rank
	Macro F1	SUBJ F1	Macro F1	SUBJ F1		
Multilingual	0.67	0.73	-	-	4/4	3/4
Arabic	0.65	0.52	-	-	5/5	4/5
Dutch	0.73	0.77	-	-	3/5	4/5
English	0.63	0.74	-	-	9/11	7/11
German	0.71	0.67	-	-	3/7	6/7
Italian	-	-	0.63	0.54	5/6	4/6
Turkish	-	-	0.70	0.79	6/6	5/6

4.2.2. Task 3A

Table 7 presents results for the validation set. We observe that zero-shot outperforms the few-shot classification. Thus, we submit zero-shot classification results for the test data. The results on the test set are shown Table 8. Our ranking in Task 3A is better than our ranking in Task 2, suggesting that ChatGPT is more suitable for political bias detection than subjectivity.

Table 7
Results for Task 3A Validation Set

Classification	**Precision**	**Recall**	F_1	**Accuracy**	**MAE**
Few-shot	0.46	0.43	0.43	0.43	0.71
Zero-shot	0.54	0.49	0.48	0.49	0.58

Table 8
Results for Task 3A Test Set

Classification	Precision	Recall	F_1	Accuracy	MAE	Rank
Zero-shot	0.49	0.44	0.40	0.44	0.64	2/5

5. Conclusion

In this paper, we present our participation in Task 2 and Task 3A of CLEF-2023 Check That! Lab. For both tasks, we utilize ChatGPT with two different approaches: zero-shot and few-shot classification.

In Task 2, we could not surpass the baseline in five languages (Multilingual, Turkish, Arabic, Italian, and English) with the approach we employed. We achieved the best results in Dutch and German languages, where we outperformed the baseline and secured the 3 position in the rankings.

for Arabic, Dutch, English, German, Italian, and Turkish languages, respectively. In Task 3A, we are ranked 2$^{\text{nd}}$ (out of 5 groups).

In the future, we plan to explore how ChatGPT can be utilized more effectively in these tasks. In particular, we plan to investigate the impact of how prompts are written, the size of the samples provided in the few-shot approach, and how to select the samples.

References

[1] Y. S. Kartal, M. Kutlu, Re-think before you share: A comprehensive study on prioritizing check-worthy claims, IEEE transactions on computational social systems 10 (2022) 362–375.

[2] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 2931–2937.

[3] A. Galassi, F. Ruggeri, A. B.-C. no, F. Alam, T. Caselli, M. Kutlu, J. M. Struss, F. Antici, M. Hasanain, J. Köhler, K. Korre, F. Leistra, A. Muti, M. Siegel, M. D. Turkmen, M. Wiegand, W. Zaghouani, Overview of the CLEF-2023 CheckThat! lab task 2 on subjectivity in news articles, in: Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum, CLEF '2023, Thessaloniki, Greece, 2023.

[4] G. Da San Martino, F. Alam, M. Hasanain, R. N. Nandi, D. Azizov, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 3 on political bias of news articles and news media, in: Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum, CLEF '2023, Thessaloniki, Greece, 2023.

[5] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struss, R. N. Nandi, G. S. Cheema, D. Azizov, P. Nakov, The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis,

C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 506–517.

[6] J. Sixto, A. Almeida, D. López-de Ipiña, An approach to subjectivity detection on twitter using the structured information, in: Computational Collective Intelligence: 8th International Conference, ICCCI 2016, Halkidiki, Greece, September 28-30, 2016. Proceedings, Part I 8, Springer, 2016, pp. 121–130.

[7] I. Chaturvedi, E. Ragusa, P. Gastaldo, R. Zunino, E. Cambria, Bayesian network based extreme learning machine for subjectivity detection, Journal of The Franklin Institute 355 (2018) 1780–1797.

[8] A. Montoyo, P. Martínez-Barco, A. Balahur, Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments, Decision Support Systems 53 (2012) 675–679.

[9] E. Riloff, Exploiting subjectivity classification to improve information extraction ellen riloff janyce wiebe william phillips (2005).

[10] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, S. Patwardhan, Opinionfinder: A system for subjectivity analysis, in: Proceedings of HLT/EMNLP 2005 Interactive Demonstrations, 2005, pp. 34–35.

[11] J. Hong, Y. Cho, J. Jung, J. Han, J. Thorne, Disentangling structure and style: Political bias detection in news by inducing document hierarchy, arXiv preprint arXiv:2304.02247 (2023).

[12] W.-F. Chen, K. Al-Khatib, H. Wachsmuth, B. Stein, Analyzing political bias and unfairness in news articles at different levels of granularity, arXiv preprint arXiv:2010.10652 (2020).

[13] Y. Liu, X. F. Zhang, D. Wegsman, N. Beauchamp, L. Wang, Politics: pretraining with same-story article comparison for ideology prediction and stance detection, arXiv preprint arXiv:2205.00619 (2022).

[14] Y. Lei, R. Huang, L. Wang, N. Beauchamp, Sentence-level media bias analysis informed by discourse structures, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 10040–10050.

[15] W.-H. Lin, T. Wilson, J. Wiebe, A. G. Hauptmann, Which side are you on? identifying perspectives at the document and sentence levels, in: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), 2006, pp. 109–116.